DRAGON SYSTEMS' 1998 BROADCAST NEWS TRANSCRIPTION SYSTEM FOR MANDARIN

Puming Zhan, Steven Wegmann, and Larry Gillick
320 Nevada Street, Newton, MA 02460

ABSTRACT

In this paper we shall describe Dragon Systems' 1998 Broadcast News transcription system for Mandarin. We shall describe our music classifier, which was unique to our Mandarin system, as well as our speaker change detection algorithm, which was used in our English and Mandarin systems. We shall also report on preliminary, post-evaluation experiments with pitch.

1. INTRODUCTION

We used the same phoneme table, pronunciation lexicon, vocabulary, and language model in our 1998 HUB4 Mandarin Broadcast News system as we used in our 1997 Mandarin Broadcast News system [1]. But we incorporated the following techniques that we developed for the 1998 English Broadcast News system [2] into our 1998 Mandarin Broadcast News system.

- We used a much more sophisticated automatic speech segmentation algorithm in the segmentation pass, which includes segmentation based on silence detection, further segmentation based on a fast Mandarin word recognizer, and refined segmentation based on speaker change detection.

- We used a diagonalizing transformation [3], instead of IMELDA, to transform the input PLP features and kept all 36 coefficients (12 PLP feature with their first and second differences).

- We trained gender dependent acoustic models for the 1998 Mandarin Broadcast news system.

- We decoded the 1998 evaluation set with three systems and combined them using ROVER [7].

In addition, we used a music detector to identify the pure music segments and discarded them in the further processing. We also explored adding pitch to our feature set in the post-evaluation experiments.

We obtained 3.1% absolute improvements on the 1997 evaluation test over our 1997 Mandarin Broadcast News evaluation system from the above new implementations.

All recognition results reported in this paper are character error rate, and they were obtained from a gender independent, no speaker normalization system with fast settings. The language model was the one used in 1997 Mandarin Broadcast News evaluation system.

2. OUTLINE OF 1998 MANDARIN BROADCAST NEWS SYSTEM

2.1 Decoding

The following is the decoding procedure we used in the 1998 Mandarin Broadcast News evaluation system:

- **Automatic segmentation:** (1) chopped the audio stream into 20 to 30 second segments based on a amplitude-based silence detection; (2) chopped these segments into 2 to 30 seconds segments based on silence obtained with a fast word recognizer; (3) refined the segments further based on the speaker change detection.

- **Music detection:** classified the segments into speech and music classes, and discarded the segments in the music class.

- **Gender detection:** a fast gender independent word recognizer was used to obtain an alignment for every segment. The alignment was scored against a small gender dependent acoustic model to determine the gender of the segment.

- **Clustering and Speaker Normalization:** all segments with same gender were clustered into several classes. Speaker normalization was performed within each cluster by doing a quick, errorful recognition with small acoustic models and a small bigram language model, and then rescoring this transcript with each warp scale in order to pick up the best scoring scale.

- **First decoding pass:** input the segments in each cluster into the system with gender dependent acoustic model and an interpolated trigram language model to obtain initial transcriptions for each cluster.

- **Second decoding pass:** One pass of unsupervised rapid adaptation with one transformation was performed within each cluster, followed by the final recognition pass using

the adapted acoustic model and the same interpolated trigram language model.

- **Using ROVER:** we ran three complete versions of this system, two with speaker normalization but using different word recognizer in the automatic segmentation pass, the other was not using the speaker normalization, and combined the outputs using ROVER [7].

With the above decoding procedure, we obtained 3.1% absolute character error reduction on the 1997 Mandarin Broadcast News evaluation set, and 20.6% character error rate on the 1998 Mandarin Broadcast News evaluation set. We got a 0.4% absolute improvement on 1997 Mandarin Broadcast news evaluation set by using ROVER.

2.2 Acoustic model

The configurations of acoustic model of our 1998 Mandarin Broadcast News evaluation system is exactly the same as our 1997 Mandarin Broadcast News evaluation system [1], except that we trained gender-dependent acoustic models using the batch adaptation algorithm. The models were trained using the same speech data as in 1997. We observed about 1.2% absolute improvement by switching to the gender dependent acoustic model. But we also observed that the gain from frequency warping speaker normalization basically disappeared, and adaptation gain was also reduced from 2.2% to 1.4% with the gender dependent model. We lost about 0.5% absolute because of mis-classification of gender. The mistakes in the automatic segmentation pass (generating some multi-speaker segments) could lead to some segments which contain multi-speaker data, hence lead to the mistakes in gender detection and estimation of the frequency warping factor. We obtained similar character error rate without using frequency warping speaker normalization. This is not surprised since the frequency warping technique basically alleviates the vocal tract difference between male and female speakers. The other big improvement we obtained in acoustic modeling was from using the diagonalizing transformation developed in Dragon's English system [3].

2.3 Language model

We did not build new language model for this year's evaluation. Therefore we just used the same language model as the one in our 1997 Mandarin Broadcast News system. This language model was the interpolated backoff trigram language model.

3. MUSIC DETECTION

3.1 Features and Model

The features we explored for music detection are:

- The features described in [4], which include 4 Hz Modulation energy, Percentage of "Low-Energy" frames, Spectral rolloff point, Spectral centroid, Delta spectrum magnitude, Zero-crossing rate, and variances of them.

- Means of 12 cepstral coefficients with their first and second differences. The cepstral coefficients were computed in a 20ms window, and their means were computed in a one-second window that contains 100 frames.

- Variances of the cepstral coefficients computed in a one-second window.

We trained a logistic regression model using these features, evaluated them on a development set, and picked up 18 most effective features among them based on the t-value in the logistic regression model training process. The features that we actually used for music detection are: Variance of spectral rolloff point, mean and variance of spectral centroid and delta spectrum magnitude, five means and eight variances of the cepstral coefficients.

3.2 Performance of music detection

The English 1995 Marketplace data with BBN's annotation of categories was used for model training and testing. We took 22.6 minutes speech data (16 minutes speech plus 6.6 minutes *music with speech*) to train the speech model, and 13 minutes pure music to train the music model. The reason to add some *music with speech* data for speech model training is to encourage this kind of data to be classified as speech, instead of music. The testing set contains 11.1 minutes pure music, 16.7 minutes speech, and 11.6 minutes *music with speech*. The training and testing speech data are gender balanced. Table 1 contains the frame accuracy of speech and music classification.

Input	S(%)	M(%)
S	92.3	7.7
M	7.4	92.6
MS	82.3	17.7

Table 1: Performance of speech and music detection counted in frames. (S = speech; M = music, MS = music with speech)

Table 1 shows that 7.7% speech frames were classified as music, 7.4% music frames were classified as speech, and 17.7% MS frames were classified as music. The results demonstrate that our music detector works well for pure speech and music classification. But the MS data is hard to classify at least in part because the ratio of speech to music is not constant.

Table 2 contains the recognition results with and without using the music detection to discard some segments. They were obtained on the Mandarin 1997 development set, and the 1997 evaluation set. The segments used in this experiment are from the reference chopping. In a reference file, a long audio stream is labeled as a series of segments. The speech segments are transcribed. Those segments which contains very noisy speech (like advertisement), overlapped speech, music, and other noises are marked as *excluded_region* or *inter_segment_gap*

which means that the output of the recognizer for these segments will not be scored. But we found that the labeled segments in a reference file usually do not cover the whole audio stream. There are some gaps in the audio stream, which are not labeled and usually are non-speech. These kind of un-labeled segments will introduce extra insertion errors if they are not excluded in recognition. Table 2 shows how much we can improve by discarding those un-labeled gaps before feeding them into the recognizer. In this experiment, we first fed all segments, including those un-labeled segments, into the recognizer, the results are in the second column of Table 2. Then we discarded those segments, which were classified as music by the music detector, and fed rest of the segments into the recognizer, the results are in the third column of Table 2. Finally, we only fed those transcribed speech segments into the recognizer, the results are in the last column of Table 2.

	MuRm	NoMuRm	Ref
97dev	18.2	19.0	18.2
97eval	21.1	21.6	19.5

Table 2: performance of music detection in recognition (NoMuRm: keep all segments; MuRm: discard the detected music segments; Ref: discard all not-scored segments)

Table 2 shows (1). The un-labeled segments did introduce extra errors; (2). Discarding some segments based on the music detection improved the performance. How much one can improve the performance by discarding some segments in this way depends on how much the un-labeled data in the input and how much music data in the un-labeled data. We also observed small improvements from using music detection in the automatic segmentation and discarding the segments detected as music. Table 2 also demonstrates the importance of removing non-speech stuff in the Broadcast News transcription task. Without classifying and removing the non-speech segments in the audio stream, the recognition error rate may be remarkably increased, though the stream be chopped exactly the same as it is manually labeled in the transcription file.

As it shows in table 1, one of the hard issues in music detection is how to distinguish the MS data from pure music. The other issue is the variants of different music. Since our music model was trained with the music in the Marketplace program, it could fail to detect some *strange* music.

4. SEGMENTATION

4.1 Segmentation Procedure

Broadcast News data comes to us in long unsegmented speech streams which not only contains speech with various speakers, backgrounds, and channels, but also contains a lot of non-speech. So it is necessary to chop the long stream into smaller segments. It is also important to make these smaller segments homogeneous (each segment only contains the data from one source) so that the non-speech can be discarded, and those segments from the same or similar source can be clustered for speaker normalization and adaptation.

In our 1997 Broadcast News Mandarin system, we produced the segments by looking for sufficiently long silence regions in the output of a coarse recognition pass. This method generated considerable multi-speaker segments, and no speaker change information was used in the segmentation.

In our 1998 Mandarin Broadcast News system, we used the speaker change detection in the segmentation pass. The following is a procedure of our automatic segmentation:

- An amplitude-based detector was used to break the input into chunks that are 20 to 30 seconds long.
- These chunks were chopped into 2 to 30 seconds long based on silences produced from a fast word recognizer.
- These segments were further refined using a speaker change detector.

4.2 Speaker change detection

Motivated by [5], we explored both the BIC algorithm described in [5], and a related, but somewhat simpler, method based on Hotelling's T^2-test [3].
Assume X(t), $0 \le t \le T$, is a feature sequence, *b* is a speaker change point, {X(t), $0 \le t \le b$}, {X(t), $b < t \le T$}, and {X(t), $0 \le t \le T$} are from Gaussian sources with μ_1, μ_2, and μ as mean and Σ_1, Σ_2, and Σ as covariance.

In the T^2-test, we used the mean, μ_1, μ_2, and covariance of whole segment, Σ, to compute the distance measure. The formulation of T^2-test is as followings:

$$T^2 = (\mu_1 - \mu_2)^T [\Sigma (1/N_1 + 1/N_2)]^{-1} (\mu_1 - \mu_2)$$

Compared to the BIC formulation in [5], variance Σ_1 and Σ_2 are not used in the T^2-test formulation. Here the T^2-test has two advantages over the BIC algorithm: (1). It does not need to re-compute the covariance while moving the search point within a window, hence it is cheap in computation; (2). We can start the search point at the very beginning of a segment, since we do not need to use the first second to initialize the variances. This makes it be able to find the change point which is close to the start and / or end of the segment, hence to handle the audio stream which has speaker changes in short period of time. We obtained similar results from these two algorithms. The common weaknesses of both algorithms are: (1). It is still a problem to find the change point in a very short period of time; (2). The algorithms are sensitive to background changes. So it could break a word because of the background change, (3). There usually exists a small bias between a detected change point and the real change point; (4) they all need to set a threshold to control the number of segments to be chopped. We

observed 10% to 15% reductions of multi-speaker segments from speaker change based segmentation.

Table 3. Contains the recognition results with the new automatic segmentation (autoSeg) and the reference-based segmentation (REF). We obtained 1.3% absolute improvement on the 1997 Mandarin evaluation test over the segmentation algorithm we used in the 1997 Mandarin system.

	AutoSeg	REF
97eval	20.0	19.5
98eval	24.6	24.6

Table 3: Performance of the automatic segmentation.

We still lose 0.5% compared to the reference-based segmentation in this set, but didn't lose on the 1998 Mandarin evaluation set. We basically observed about 0.5% loss with the automatic segmentation on the English evaluation sets (96eval, 97eval and the second set of 98eval). But we lost 1.3% on the first set of 1998 evaluation set, because of the noisy and very short speaker turns in the stream. This means that the performance of the current segmentation algorithm depends on the quality of the speech data. It could be worse if there are frequent speaker changes in short period of time in the audio stream.

5. USING PITCH

We did distinguish tones in our acoustic models, but did not use a pitch feature in the 1998 Mandarin evaluation system. Therefore, we tried to include pitch as a feature in the post evaluation experiments. The way we computed and used pitch was based on Dragon Systems' 1997 CallHome Mandarin system [6]. Pitch was detected using a spectrally-flattened autocorrelation algorithm. The normalized pitch and delta pitch were combined with the standard 36 PLP coefficients, and the feature number was reduced to 24 after IMELDA transformation. Table 4 contains the results of using pitch as an extra feature in the system.

	Without Pitch	With Pitch
97dev	18.7	18.4
97eval	20.1	19.7
98eval	27.5	27.1

Table 4: Performance of using pitch as feature.

Table 4 shows that we got about a 0.4% absolute improvement by adding pitch and delta pitch to our feature set.

The improvement is similar as it was observed in the Mandarin Callhome system [6]. But we found that the pitch features have somewhat weird distribution. So maybe they were not properly normalized. We need to investigate it in the future.

6. CONCLUSION

We have made significant improvement for our 1998 Mandarin Broadcast News system in the preprocessing and acoustic modeling. For Mandarin, it makes more sense to use sub-syllable or syllable than phoneme as speech unit. The pitch does not seem to help much in our experiments. But it should carry the tone information that is very important in Mandarin. So there might be a better way to use pitch. Different ways of word segmentation will affect the language model, hence the performance of the entire system. This is also interesting to explore.

ACKNOWLEDGEMENTS

This work was supported by the Defense Advanced Research Projects Agency. The views and findings contained in this material are those of the authors and do not necessarily reflect the position or policy of the U.S. Government and no official endorsement should be inferred.

REFERENCES

[1] P. Zhan, S. Wegmann, and S. Lowe "Dragon Systems' 1997 Mandarin Broadcast News System", *Proceedings of the Broadcast News transcription and Understanding Workshop*, Lansdowne, VA, pp. 25-27, Feb. 1998.

[2] S. Wegmann, P. Zhan, I. Carp, M. Newman, J. Yamron, and L. Gillick "Dragon Systems' 1998 Broadcast News Transcription System", This *Proceedings*.

[3] S. Wegmann, P. Zhan, and L. Gillick "Progress in Broadcast News Transcription at Dragon Systems", *To appear in Proceedings of ICASSP'99*, Phoenix, Arizona, March 1999.

[4] E. Scheier and M. Slaney "Construction and Evaluation of A Multifeature Speech/Music Discriminator", *Proceedings of ICASSP'97*, pp. 1331 – 1334, Munich, Germany, 1997.

[5] S. Chen and P. Gopalakrishnan "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion", *Proceedings of the Broadcast News transcription and Understanding Workshop*, Lansdowne, VA, pp. 127-132, Feb. 1998.

[6] Y. Ito et al., "Dragon Systems' 1997 Mandarin CallHome Evaluation System", Proceedings of HUB-5 Conversational Speech Recognition Workshop, MITAGS, November 1997.

[7] J. Fiscus, "A post-processing System to Yield Reduced Error Rates: ROVER", *Proceedings of the 1997 IEEE ASRU Workshop*, Santa Barbara, December 1997.